Supplementary Information 2: Details to background, sample and methods

On the non-linearity of radon-induced lung cancer

Environmental Science and Pollution Research

Albert Rosenberger, Heike Bickeböller, David C Christiani, Geoffrey Liu, Matthew B Schabath, Luisa F. Duarte, Loic Le Marchand ,Christopher Haiman, Teresa Landi, Dario Consonni, John K Field, Michael P.A. Davies, Demetrios Albanes, Adonina Tardon, Guillermo Fernández-Tardón, Gad Rennert, Hedy Rennert, Christopher I Amos, Rayjean J Hung

Corresponding Author

Albert Rosenberger Universitätsmedizin Göttingen, Institut für Genetische Epidemiologie, Göttingen, Germany, Email: arosenb@gwdg.de

1 Overview: results of previously meta-analyses and systematic reviews

S2-Table 1 Excess odds ratio (EOR) estimates of previously meta-analyses and systematic reviews

	Author	Year	Included studies	constrains	EOR per 100 Bq/m³	conclusion
i	Darby et al.	2005	13 case-control studies	Europe	8.4% (95%-CI: 3.0% to 15.8%)	LNT
ii	Krewski et al.	2006	7 case-control studies	North-American	10% (95% CI: -1% - 26%)	no evidence of deviation from LNT
iii	Lubin et al.	2004	2 case-control studies	China	13.3% (95%-CI: 1%-36%)	LNT>80 Bq/m³
iv	Malinovsky et al.	2019	31 case–control studies,	1990-2014	14% (95% CI: 8%–21%)	LNT >100 Bq/m³
v	Cheng et al.	2021	pooled estimates	never smokers	15% (95% CI 6%–25%)	
				ever smokers	9% (95% CI 3%–16%).	

EOR excess odds ratio per 100 Bq/m³, LNT Linear non-threshold model

i. Darby et al. (2005): The association of residential radon exposure and lung cancer was determined by a collaborative analysis of individual data from 13 European case-control studies of residential radon and lung cancer (in total 7.148 cases and 14.208 controls). The risk of lung cancer increased by 8.4% (95%-CI: 3.0% to 15.8%; P = 0.0007) per 100 Bq/m³ (measure radon) was estimated. A linear dose-response relation with no evidence of a threshold was seen. No evidence was found that the risk per 100 Bq/m³ differed according to age, sex, or smoking.

ii. Krewski et al. (2006) summarizing data of 4081 cases and 5281 controls collected in 7 case-control studies from North-America. The estimated excess OR (EOR=0.10 per Bq/m³ (95% CI: -0.01 - 0.26) for time-weighted radon concentration in a 5–30 years interval prior to the index date. They report no evidence of a deviation from linearity in the EOR.

iii. Lubin et al. (2004) summarize 2 case-control studies conducted in China with 1,050 lung cancer cases and 1,996 controls. Based on a linear model, the EOR per 100 Bq/m³ was 13.3% (95%-CI: 1%-36%). However, the model presented (Figure 1a) indicate an OR>1 (EOR>0) for radon concentrations *above ~80 Bq/m³*.

iv. Malinovsky et al. (2019) summarizes 31 case–control studies (published 1990-2014) with in total 20,703 cases, 34,518 controls and 140 individual odds ratio (OR) estimates. They found a *linear no-threshold exposure–effect relationship* fitting best (compared to a model with threshold) for radon concentrations *above 100 Bq/m³ (*lower threshold*)*, with a slope of 14% (95% CI: 8%–21%) per 100 Bq/m³. (Encountering the hypothesis of Dobrzyński et al. (2018): Radon causes no statistically significant increase in lung cancer incidence at radon concentrations below 1000 Bq/m³).

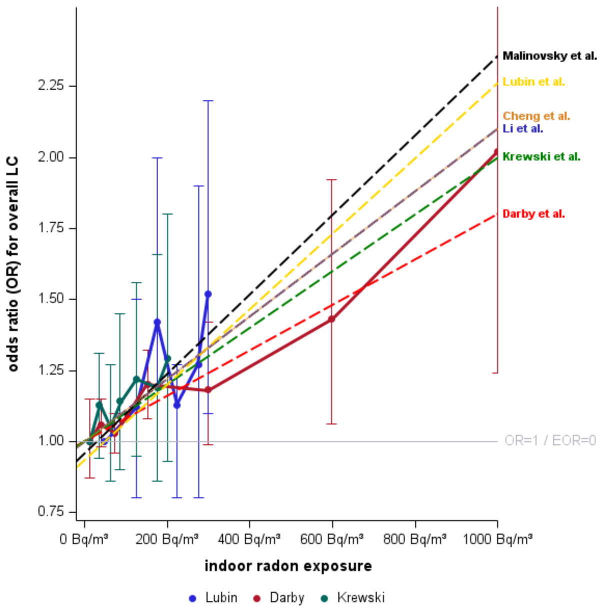
v. Cheng et al. (2021) performed a meta-analysis on four pooled estimates: (Europe) Darby et al. (2005), (Noth America) Krewski et al. (2006), (China) Lubin et al. (2004) and (Northwest Spain) Lorenzo–Gonzalez et al. (2019). Information of 2341 never smoker cases, 8967 never smoker controls, 9937 ever-smoker cases and 12 463 ever-smoker controls entered the considerations. The estimated an EOR per 100 Bq/m³ for never smokers of 15% (95% CI 6%–25%) and for ever smokers of 9% (95% CI 3%–16%). Although these estimates were not statically significant (p=0.32), the difference is remarkable large. However, they found the EOR within never smokers significantly modified by sex: women: 9% (95% CI −2%–20%); men: 46% (95% CI 15%–76%); p=0.027.

S2-Table 2 Effect estimates (OR and EOR) for indoor radon

	***Li et al.* (2020) (incl. Darby, Krewski, Lubin)**				***Krewski et al.* (2005)** *all LC-cases**		**2-20 years radon measurement ** **		***Darby et al.* (2006)**	
	OR$	*95%-CI*	***EOR***	*95%-CI*	***EOR***	*95%-CI*	***EOR***	*95%-CI*	***EOR***	*95%-CI*
SCLC	*2.03*	*1.52–2.71*	***19%***	*7%–32%*	***0.23***	*-0.08–0.85*	***0.20***	*-0.11–1.00*	***31%***	***12%-61%***
AdenoLC	*1.58*	*1.31–1.91*	***13%***	*1%–25%*	***0.09***	*-0.05–0.33*	***0.27***	*0.02–0.73*	*6%*	*-3% - 20%*
SqCLC	*1.43*	*1.18–1.74*	***--***	*-12%–12%*	***0.05***	*-0.04–0.33*	***0.13***	*-0.04-0.62*	*-1%*	*-3% - 9%*
others	*1.54*	*1.11–2.15*	***3%***	*-22%–16%*	***0.22***	*-0.04–0.84*	***0.22***	*-0.04-0.84*	*3%*	*-3% - 24%*
Europe	*1.77*	*1.54–2.03*							***8%***	*3%-16%*
North America	*1.09*	*0.94–1.27*			***0.10***	*-0.01–0.26*	***0.18***	*0.02–0.43*		
Asia	*0.93*	*0.42–2.06*								
overall	*1.48*	*1.26–1.73*	***11%***	*5%–17%*	***0.10***	*-0.01–0.26*	***0.18***	*0.02–0.43*	***8%***	*3%-16%*

(Text and Table 2 in Li et al. 2020) or (Table 11a/12a in Krewski et al. 2005)) or (Text and Tabelle 24 in Darby et al. 2006); $ „fixed effects“ OR estimator for the indoor radon association; * “5–30 yr exposure time window”** „subjects residing in 1 or 2 houses during the exposure window and with 20 yr or more covered with alpha-track air monitors“; EOR... excess odds ratio je 100 Bq/m³

S2-Figure 1 published LNT models



2 ILCCO Studies (which contributed data to this study)

S2-Table 3 Data contribution ILCCO studies

Study Name (lang)	Study acronym	PI	Country	Year	Control eligibility/ recruitment	Matching factors
Harvard Lung Cancer Study	HLCS	D.C. Christiani	US	1992-2004	friends/spouse	none
Mount-Sinai Hospital-Princess Margaret Study	MSH-PMH	R. Hung / G. Liu	Canada	2008-2012	patients registered at the family medicine clinic	Age, sex
International Early Lung Cancer Action Program	IELCAP-Toronto	G. Liu	Canada	2004-2011	>10 packyear, >50 yrs	Age, sex, smoking
Total Lung Cancer study	TLC	M. Schabath	US	2012-Present	friends, Lifetime screening clinic, Lung and thoracic tumour education, patient advocacy program	age, gender, ethnicity and smoking
Multiethnic Cohort	MEC	L. LeMarchand	US	1993-1996	45-74 yrs, holder of a driver license	prospective follow up
Malmö Diet and Cancer Study	MDCS	H. Brunnström	Sweden	1991-present	44-74 yrs, living in Malmö	prospective follow up
Environment And Genetics in Lung cancer Etiology	EAGLE	M.T. Landi	Italy	2002-2005	health registry	Age, sex
Liverpool Lung Project	LLP	J.K. Field	UK	1999-2016	resident, no previous cancer	Age, sex
Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study	ATBC	D. Albanes	Finland	1985-1993	male smokers, 50-69-yrs	prospective follow up
German Lung Cancer Study	GLCS	E. Wichmann, A. Risch, H. Bickeböller	Germany	2000-2004	age < 50, KORA	residence (southern Germany)
CAncer de PUlmon en Asturias	CAPUA	A. Tardon	Spain	2002-2012	resident, no previous cancer	sex
Israel Lung Cancer Study	NICCC-LCA	G. Rennert	Israel	2005-present	Population registry	Age, sex, clinic

Excluded from this data analysis: ATBC (Finland), LLP (UK), GLCS (Germany), MDCS (Malmö, Sweden), MEC (Hawaii, USA) and TLC Moffitt (USA);
included in this data analysis: CAPUA (Spain), EAGLE (Italy), HLCS (US), NICCC-LCA (Israel), MSH/PMH (Canada)

S2-Table 4 Representative surrounding area of assigned sIR

surrounding area	total		CAPUA	EAGLE	HLCS	NICCC-LCA	MSH-PMH
total	14,489	100%	1,683	3,856	5,858	1,184	1,908
FSA 1-digit	832	6%	--	--	--	--	832
FSA 2-digit	853	6%	--	--	--	--	853
FSA 3-digit	223	2%	--	--	--	--	223
city	280	2%	--	--	280	--	--
US county	5,578	38%	--	--	5,578	--	--
grid cell	1,683	12%	1,683	--	--	--	--
municipality	5,026	35%	--	3,842	--	1,184	--
province	5	<1%	--	5	--	--	--
region	9	<1%	--	9	--	--	--

FSA forward sortation area (geographical region[9] of Canadian postal codes; grid cell 10 km x 10 km grid cells of the European Indoor Radon Map[12]; province/region 1^{st}- and 2^{nd}-level constituent entities of Italy

S2-Table 5 Mean indoor radon exposure in the surrounding area [Bq/m³] – continuous

sIR: continuous	n	mean	median	range	std. dev.
total	14,489	89.6	71.1	4.0-835.0	49.02
CAPUA	1,683	94.6	89.3	13.0-287.2	29.69
EAGLE	3,856	64.1	64.3	21.2-249.2	20.84
HSPH	5,858	122.4	152.0	4.0-835.0	54.72
NICCC-LCA	1,184	45.3	42.9	31.8-79.0	7.14
MSH-PMH	1,908	63.5	50.3	27.5-258.7	28.84
histological subtype					
controls	5,562	84.3	64.8	7.0-835.0	46.67
cases	8,927	92.9	74.0	4.0-835.0	50.15
unknown	689	97.5	74.0	15.0-329.0	54.00
SqCLC	2,210	91.4	74.0	4.0-835.0	50.22
SCLC	730	96.9	83.9	15.0-329.0	45.06
LCLC	408	96.8	78.0	19.0-318.0	45.94
AdenoLC	4,890	92.0	74.0	4.0-835.0	50.55

sIR mean indoor radon exposure in the surrounding area [Bq/m³]; std. dev. standard deviation; LCLC: large cell lung cancer, SCLC: small cell lung cancer, SqCLC: squamous cell lung cancer, AdenoLC: lung adenocarcinoma

S2-Table 6 Mean indoor radon exposure in the surrounding area [Bq/m³] - categorized

sIR: grouped	n	<25 Bq/m³		25-<50 Bq/m³		50-<75 Bq/m³		75-<100 Bq/m³		100-<150 Bq/m³		150-<200 Bq/m³		200- Bq/m³	
total	14,489	131	<1%	2,357	16%	5,366	37%	1,953	13%	1,630	11%	2,825	19%	227	2%
CAPUA	1,683	6	<1%	105	6%	216	13%	966	57%	385	23%	--	--	5	<1%
EAGLE	3,856	4	<1%	696	18%	2,633	68%	350	9%	133	3%	34	<1%	6	<1%
HSPH	5,858	121	2%	90	2%	1,292	22%	445	8%	952	16%	2,750	47%	208	4%
NICCC-LCA	1,184	--	--	1,069	90%	114	10%	1	<1%	--	--	--	--	--	--
MSH-PMH	1,908	--	--	397	21%	1,111	58%	191	10%	160	8%	41	2%	8	<1%
histological subtype															
controls	5,562	48	<1%	1,055	19%	2,195	39%	811	15%	533	10%	844	15%	76	1%
cases	8,927	83	<1%	1,302	15%	3,171	36%	1,142	13%	1,097	12%	1,981	22%	151	2%
unknown	689	10	1%	112	16%	227	33%	68	10%	78	11%	170	25%	24	3%
SqCLC	2,210	12	<1%	313	14%	795	36%	364	16%	275	12%	413	19%	38	2%
SCLC	730	5	<1%	77	11%	253	35%	119	16%	91	12%	175	24%	10	1%
LCLC	408	4	<1%	46	11%	150	37%	31	8%	56	14%	118	29%	3	<1%
AdenoLC	4,890	52	1%	754	15%	1,746	36%	560	11%	597	12%	1,105	23%	76	2%

sIR mean indoor radon exposure in the surrounding area [Bq/m³]; LCLC: large cell lung cancer, SCLC: small cell lung cancer, SqCLC: squamous cell lung cancer, AdenoLC: lung adenocarcinoma

3 Spatial indoor radon exposure (sIR)

S2-Table 7 Source or sIR and assignment to study participants

	provider of sIR data	data source	ref.	data grid	survey years	linked LC-study	recruitment year	address link to sIR data
AUSTRIA	Institut für Isotopenforschung und Kernphysik der Universität Wien	electronic report	(Friedman 2007; European Commission et al. 2019)	counties (Gemeinde)	1992-2003	GLCS	2000-2004	by geo-coordinates
CANADA	Radiation Health Assessment Division (RHAD), Health Canada - Radiation Protection Bureau	electronic record [$]	(2012)	counties and 3-diget post codes	2009-2011	MSH-PMH	2008-2012	by 3-diget postal codes
DENMARK	Centre for Nuclear Technologies of the Technical University of Denmark (DTU)	electronic report	(Andersen et al. 2001; European Commission et al. 2019)	counties (Kommunenavn)	1995-1996	not linked	--	by county names
FINLAND	Radiation and Nuclear Safety Authority (STUK)	electronic record	(Valmari et al. 2010; Tollefsen et al. 2014; European Commission et al. 2019)	10x10 km grid	1986-2004	ATBC	1985-1993	by geo-coordinates
GERMANY	Bundesamt für Strahlenschutz	electronic record	(Tollefsen et al. 2014; European Commission et al. 2019)	10x10 km grid	1978-2003	GLCS	2000-2004	by geo-coordinates
ISRAEL	Soreq Nuclear Research Center	electronic report (Hebrew version)	(G. Haquin et al. 2012)	counties (שם האזור)	1998-2012	NICCC-LCA	2005-	by Natural Region
ITALY	Centro Nazionale per la Protezione dalle Radiazioni e Fisica Computazionale	electronic record	(Tollefsen et al. 2014; European Commission et al. 2019)	counties (comune)	1989-1998 2004-2007	EAGLE	2002-2005	by county names
SPAIN	---	European Indoor radon map	(Tollefsen et al. 2014; European Commission et al. 2019)	10x10 km grid	1989-2010	CAPUA	2002-2012	by geo-coordinates
SWEDEN	Boverket – the Swedish National Board of Housing, Building and Planning	electronic record	(European Commission et al. 2019)	counties (commune)	1991-1992	MDCS	1991-	by county names
UNITED KINGDOM	Radon Group, Public Health England (PHE)	electronic record	(Grainger et al. 2000; European Commission et al. 2019)	10x10 km grid	1980-2005	LLP	1999-2016	via postal codes and geo-coordinates [$$]
UNITED STATES OF AMERICA	US Environmental Protection Agency (EPA)	EPA's state radon map book [§]	(1993)	counties / cities	1986-1992	HLCS MEC TLC	1992-2004 1993-1996 2012-	by county names [&&] equal sIR values [&] by county names [&&]

sIR spatial indoor radon; ref. reference, geolocation: A tool for converting geo-coordinates coordinates into grid cells of the European indoor radon map was provided to us by P. Bossew from the Federal Office for Radiation Protection, Berlin; [&] Since there is almost no radon exposure in Hawaii, all participants were assigned the same value. [&&] taking the US states into account. [$$] postal codes were transformed into geo-coordinates using https://www.doogal.co.uk/UKPostcodes.php; [$] publically available at https://open.canada.ca/data/en/dataset/744d8a3b-b9e0-41b8-be5f-5f869a36a221; [§] publically available at https://nepis.epa.gov/; Study name abbreviations: see S2-Table 3.

4 Spatial indoor radon exposure in the surrounding area: main response variable sIR

S2-Table 8 Missing summary measures for sIR in the surrounding area

Indoor radon of surrounding area		mean		median		geom.mean		mean of ln		US radon zone	
	N	NMiss	%	NMiss	%	NMiss	%	NMiss	%	NMiss	%
ATBC	569	--	--	--	--	569	100%	569	100%	569	100%
CAPUA	718	2	0%	2	0%	2	0%	2	0%	718	100%
EAGLE	1.618	30	2%	30	2%	1618	100%	1618	100%	1618	100%
LLP	411	258	63%	258	63%	411	100%	411	100%	411	100%
GLCS	23	--	--	2	9%	14	61%	14	61%	23	100%
HSPH	2.295	11	0%	253	11%	213	9%	2295	100%	124	5%
NICCC-LCA	583	2	0%	583	100%	583	100%	583	100%	583	100%
MDCS	78	--	--	--	--	--	--	--	--	78	100%
MEC	177	--	--	--	--	--	--	177	100%	177	100%
TLC	390	5	1%	384	98%	384	98%	390	100%	5	1%
MSH-PMH	1.394	17	1%	1394	100%	1394	100%	1394	100%	1394	100%
total	***8.256***	*325*	*4%*	*2906*	*35%*	*5188*	*63%*	*7453*	*90%*	*5700*	*69%*

NMiss number of missing values; geom.mean geometric mean, US United States of America

S2-Table 9 Correlation of summary measures for sIR in the surrounding area

	mean	median	geom.mean	mean of ln(IR)
mean	Person→	0.90279 <.0001 5350	0.82713 <.0001 3068	0.91310 <.0001 803
median	0.85858 <.0001 5350		0.98609 <.0001 3028	0.96900 <.0001 803
geom. mean	0.77379 <.0001 3068	0.97850 <.0001 3028		0.96608 <.0001 803
mean of ln(IR)	0.71460 <.0001 803	0.98737 <.0001 803	0.98746 <.0001 803	←Spearman

Top-right: Pearson correlation; bottom-left: Spearman rank correlation; presented are also p-value / number of pairs observed

5 Spatial indoor radon exposure (sIR): Extreme values

Extreme values (outliers) of sIR were identified and checked for correctness. This refers to values fulfilling $|r| \geq 5s$ (r residual, s robust standard error) after applying Optimal Locally Robust M-Estimates (method MM) as implemented in PROC ROBUSTREG (SAS®). Such sIR values may represent unrealistic proxies of individual exposure, due to the spatial sampling process within dwellings, and may introduce bias during parameter estimation. Thus, we identified and replaced such extreme values, as recommended for constructing radon risk maps (Murphy and Organo 2008).

To **largest 75 sIR** were considered as extreme value, all **exceeding** a threshold of **329 Bq/m³**. Extreme sIR values in this order of magnitude were assigned to 63 participants of ATBC (Finland, 329-835 Bq/m³), 10 of HSPH (329-835 Bq/m³) and one each of TLC (396 Bq/m³) and GLCS (337 Bq/m³). Note, that the corresponding sIR-median was more than 100 Bq/m³ smaller. All extreme values belong to geographical regions, known to be radon prone: County Carroll, New Hampshire; and regions of Lahti, Kotka and Kouvola in Finland. These values were replaced by the respective quantiles of a log-normal-distribution[$] for more robust modelling. This replacement is less restrict than trimming or winsorizing of outliers.

[$] Parameters were estimated by fitting a truncated log-normal -distribution to all non-extreme values (see chapter 6 Spatial indoor radon exposure (sIR): Distribution).

6 Spatial indoor radon exposure (sIR): Distribution

sIR values of representative dwelling samples are known to follow a skewed and heavy tailed distribution. The fitting of inter alia lognormal or gamma distributions to spatial indoor radon values have been discussed[a] (Tuia and Kanevski 2008; Murphy and Organo 2008; Bossew 2010; Daraktchieva et al. 2014). sIR values can only be assumed as roughly log-normal distributed within windows up to a few 10 km radius (and under some other conditions) and after outliers or extremes have been removed, but having a somewhat fatter tail than expected for log-normality (Bossew 2010).

The sample at hand is mixture of not representative LC cases and controls from several study sites, with most catchment areas by far larger than 10 km radius. Hence, an appropriate right-skewed fat-tail distribution needed to be determined for data modelling.

We compared the fit of several truncated candidate-distributions (listed in S2-Table 10) to the observed sIR values by the Kolmogorov-Smirnov Statistic as implemented in PROC SEVERITY of SAS® (in addition to AIC and BIC). The **log-normal distribution** (parameters) to **fit best** to the **observed sIR** values. The fit of a Burr (Type XII) and the gamma-distribution are comparable, even preferable according AIC/BIC criteria. However, the Burr distribution cannot be used fitting linear mixed model (used for data modelling) (see S2-Figure 2).

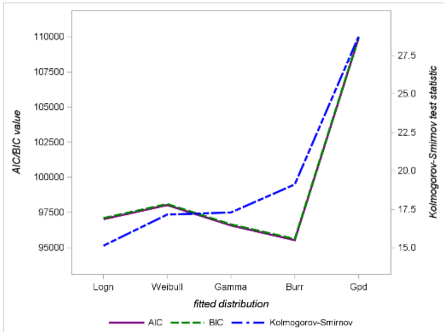
To avoid miss fitting due to outliers, we considered truncation at a threshold of **259 Bq/m³** (see chapter *5* Spatial indoor radon exposure (sIR): Extreme values).

S2-Table 10 Fit of heavy tail distribution to sIR-values

Distribution	converted	KS	selected
Gamma	yes	17.27470	no
Burr	yes	19.12046	no
Weibull	yes	17.19577	no
Logn	yes	**15.10954**	**yes**
Gpd	yes	28.74406	no

KS Kolmogorov-Smirnov Statistic

S2-Figure 2 Fit of heavy tail distribution to sIR-values



[a] *"The term 'log-normal mysticism' has been coined by Tóth et al. (2006) for the so far unexplained observation that indoor radon concentration data from geographical surveys very often appear to be quite accurately log-normally distributed."* (Bossew 2010)

7 Definition of parametric models

The logistic model is defined as:

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_r r+\boldsymbol{\beta X}} \qquad \text{(Equation 1)}$$

where π is the probability being a lung cancer case in the sample and $\frac{\pi}{1-\pi}$ the corresponding odd; r is the exposure (here sIR) and X a matrix of covariates. β_0, β_r and **β** are model parameter that need to be estimated. According to the theory of case-control studies one can approximate the relative risk of lung cancer in term of the ratio of odds of an exposed to an unexposed individual, resulting in $OR = e^{\beta_0}$

The linear non-threshold model (LNT) is defined as:

$$\frac{\pi}{1-\pi} = (1+\beta_r r)e^{\beta_0+\boldsymbol{\beta X}} \qquad \text{(Equation 2)}$$

From this definition follows, that the odds ratio is $OR = (1+\beta_r r)$, where β_r is referred as excess odds ratio (EOR). β_r can be interpreted as the percentage change of OR by one unit of the exposure r. $e^{\beta_0+\boldsymbol{\beta X}}$ can be considered as baseline odd.

Per definition, a continuous increase (decrease) in the risk is modelled the further the exposure r is different from 0. An exposure of r=0 (even if not exists) is considered as the "lowest (largest) risk exposure level" (LRE).

We also fitted a modified version of LNT (referred as LNT+), incorporating a shift parameter β_t.

The linear shifted non-threshold model (LNT+) is defined as:

$$\frac{\pi}{1-\pi} = (1+\beta_r(r-\beta_t))e^{\beta_0+\boldsymbol{\beta X}} \qquad \text{(Equation 3)}$$

The reference exposure level (REL) corresponding to the baseline odd (respectively an OR=1) is then $r-\beta_t$.

The linear threshold model (LT) is defined as:

$$\frac{\pi}{1-\pi} = \begin{matrix} (1+\beta_r(r-\beta_t))e^{\beta_0+\boldsymbol{\beta X}} & r > \beta_t \\ e^{\beta_0+\boldsymbol{\beta X}} & r \le \beta_t \end{matrix} \qquad \text{(Equation 4)}$$

In contrast to the LNT, the lowest (largest) risk is assigned to every individual with exposure r$\le \beta_t$; a continuous increase (decrease) in the risk is modelled the further the exposure r exceeds β_t.

The linear mirror point model (further referred as LT+) is defined as:

$$\frac{\pi}{1-\pi} = \begin{matrix} (1+\beta_r(r-\beta_t))e^{\beta_0+\boldsymbol{\beta X}} & r > \beta_t \\ e^{\beta_0+\boldsymbol{\beta X}} & r = \beta_t \\ (1+\beta_r(\beta_t-r))e^{\beta_0+\boldsymbol{\beta X}} & r < \beta_t \end{matrix} \qquad \text{(Equation 5)}$$

Here, the LRE is considered to be at β_t, with continuous increase (decrease) in the risk the further the exposure r differs from β_t.

8 Spline models

To overcome the limitation of the linearity assumption of the parametric models, we fit logistic regression model with spline functions for sIR. "A spline function is a piecewise polynomial function in which the individual polynomials have the same degree and connect smoothly at join points whose abscissa values, referred to as [internal] knots, are pre-specified." (SAS 2013) The degree (dg) of these piecewise polynomial function need to be set in advanced, e.g. dg=0 (to construct step functions), dg=1 (lines connect the knots), dg=2 (piecewise quadratic curve; where slopes coincide at the knots) or dg=3 (piecewise cubic curve); where slopes and curvature coincide at the knots). SAS PROC LOGISTIC provides the fit of B-splines, where additional (so-called boundary) knots are placed automatically outside the range of pre-specified knots. We also fitted natural cubic splines, also known as "restricted cubic splines", where the function is linear beyond the extreme knots. It is known that the choice of the pre-defined internal knots and the "spline basis" change the model fit (Harrell 2001).

Spline settings

Therefore, we defined M=44 settings to estimate spline functions, differing by internal knots and the "spline basis". We positioned internal knots either a) by typical radon thresholds (e.g. 50, 100 or 200 Bq/m³), b) as "equally spaced knots" in terms of exposure quantiles (e.g. then 25%-, 50%- and 75%-quartiles) or c) according to a recommendation by Harrell et al. (Harrell 2001). 'We further fit B-splines (Bsp) and natural cubic splines (NCS). A list of all spline definition is given in (see S2-Table 11).

Modelling strategy

We next applied the following strategy, which results in model averaging.

i. *Fitting splines*: We applied up to $M = 44$ different spline definitions (*sp* for model m) within the framework of logistic regression, varying by internal knots and spline basis (see S2-Table 11) and one basic model without a spline for sIR: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + sp_m(sIR) + \boldsymbol{\beta X}$
with π the case-probability within the sample, sp_m the spline-function for sIR, and $\boldsymbol{\beta X}$ to adjust for other covariates.
The model (spline) fit was assess by a range of Bayesian-type generalized information criteria $BGIC(c,r)$ (i indexing a set (c,r)); including the specification of Akaike information criterion AIC=BGIC(c=2, r=0) and Bayesian information criterion BIC=BGIC(c=1, r=1)).

ii. *Deriving weights*: In order to average over all splines $sp_m(sIR)$, instead of lasting on a single model, we searched for weights $w_{m,i}$ from those $BGIC(c,r)$, which provides the lowest concentration (quantified by Hirschman-Herfindahl-Index HHI_i. Details see below.
The ΔAIC (difference in AIC of models with and without a spline for sIR) was found to be almost optimal for model weighting for the overall LC as well as for the considered subgroups. Hence, ΔAIC-based results are reported.

iii. *Lowest risk exposure level (LRE)*: The sIR at the minimum case-probability of the estimated spline function $sp_m(sIR)$ (if some exist between 15 and 120 Bq/m³) is consider as the *lowest risk exposure level* (*LRE$_m$*) and were averaged over all models. This weighted mean is presented as LRE along with the 95% confidence intervals and the range. This was applied to the total sample (for overall LC) as in subgroups.

iv. *Model specific odds ratios*: $OR_{m,sIR}$ estimates were derived for each "spline model" m at pre-defined exposure levels (e.g. sIR=100 Bq/m³), relative to the reference exposer of LRE at 58 Bq/m³.

v. *Average odds ratios*: We finally averaged the corresponding $OR_{m,sIR}$ estimates to $\overline{OR}_{sIR}$ applying a weighted version Rubin's equation (Rubin 2004; Wang et al. 2009). Hence, in constructing 95% confidence intervals for OR_{sIR} we took into account estimating and modelling uncertainty.

We fitted the spline functions adjusted for sex, age and smoking (type: never, former and current smoker; age at smoking initiation; time since stop smoking; pack years). We repeated the analysis in subgroups by sex, age (split into 4 groups defined by quartiles), smoking type (never smokers; former and current smokers split into thirds by pack years), participating study, and by histological subtype (SCLC, SqCLC, LCLC and AdenoLC).

Deriving Odds ratios from spline function

According to the logistic regression framework, we calculated the odds ratio at exposure level sIR relative to a reference value at 64 Bq/m³ (median in the total sample), as $OR_{m,sIR} = \frac{odd_{m,sIR}}{odd_{m.64\,Bq/m^3}} = e^{sp_m(sIR)-sp_m(64)}$, with $sp_m(sIR)$ the spline-function *m*.

Definition of BGIC, with AIC and BIC as special case

The *B*ayesian-type *G*eneralized *I*nformation *C*riteria is defined as $BGIC(c,r) = -D(\theta) + c \cdot d \cdot (logN)^r$ (Lu 2008), where $D(\theta)$ is the model deviance ($D(\theta) = 2\log(p(\boldsymbol{y}|\theta))$), d is the number of free parameters in the model and N is the sample size. Setting c=2 and r=0 results in $BGIC = AIC = -D(\theta) + 2d$. Setting c=1 and r=1 results in $BGIC = BIC = -D(\theta) + d(logN)$. We calculated $BGIC(c,r)$ for r= 0 to 2 by 0.5 and c=0 to 3 by 0.5, differently weighting the number of free parameters, respectively the sample size.

Deriving model weights $w_{m,i}$

We estimate model weights (as recommended (Wang et al. 2009)) by $\widehat{w}_{m,i} = \frac{e^{-\Delta BGIC_{m,i}/2}}{\sum_{k=1}^{M} e^{-\Delta BGIC_{k,i}/2}}$, where M is the number of "spline models", and - to correct for the model fit attributable to covariates - $\Delta BGIC_{m,i} = BGIC_{basis,i} - BGIC_{m,i}$ with $BGIC_{basis}(c,r)$ the corresponding information criteria of the model without spline for sIR. Alternative weights were calculated without correction by $BGIC_{basis,i}$.

Concentration of model weights $w_{m,i}$: Herfindahl-Hirschman Index

For each $BGIC(c,r)$ we calculated the Herfindahl-Hirschman Index $HHI_i = \frac{\sum_{m=1}^{M} w_{i,m}^2 - 1/M}{1-1/M}$, (Bleymüller et al. 2008) as concentration measure for distributions of discrete random variables with M possible realisations. HHI ranges between 0 (all $w_{i,m} = \frac{1}{M}$, minimal concentration) and 1 (one $w_{i,m} = 1$ all other 0, maximal concentration).

Weighted Rubin' equation

We applied a weighted adaption(Burnham and Anderson 2002) of Rubin method(Rubin 2004) on pooling (aver-aging) estimates over multiple samples for on the same date on the odds ratio estimate

Consider θ_m to represent the Odds ratio estimate $OR_{m,sIR}$ derived from spline *m* at an exposure level of *sIR*.

The model-averaged value $\bar{\theta}$ can be estimated as under the condition of $\sum_{m=1}^{M} \widehat{w}_m = M$.

$$\hat{\hat{\theta}} = \sum_{m=1}^{M} \widehat{w}_m \hat{\theta}_m,$$

The sampling variance of the estimator of $\bar{\theta}$ can be derived from the between sample variance (representing model uncertainty) and then within sample variance (representing estimation uncertainty).

The between sample variance B can be estimated as $B = \frac{1}{M} \sum_{m=1}^{M} \widehat{w}_m (\hat{\theta}_m - \hat{\hat{\theta}}_m)^2.$

Then average within sample variance V can be calculated as $\bar{V} = \sum_{m=1}^{M} \widehat{w}_m \underline{s.e(\hat{\theta}_m)^2}.$

The variance estimate associated with $\hat{\hat{\theta}}$ is the total variance $T = V + \left(1 + \frac{1}{M}\right) B.$

The statistic $(\theta - \hat{\hat{\theta}}) T^{-\frac{1}{2}}$ T follows approximately $\sim t_{v_m}$, with $v_m = (M-1)\left[1 + \frac{\bar{V}}{\left(1+\frac{1}{M}\right)B}\right]^2$ degrees of freedom.

S2-Table 11 Spline definition and BGIC based model weights for overall LC

spline definition				model rank				model weight:	
basis	internal knots	degree	knot setting	AIC	BIC	ΔAIC	ΔBIC	best ΔBGIC(c=1, r=0.5)	worst BGIC35
Bsp	24 44.2 54.4 65.1 80.1	2	equally spaced	1	18	25%	0%	17%	0%
Bsp	44.2 54.4 65.1 80.1	2	equally spaced	2	16	11%	1%	10%	0%
Bsp	37 64 151	1	Rn-values	3	1	10%	44%	15%	0%
Bsp	37 64 100 151	1	Rn-values	4	5	7%	5%	8%	0%
Bsp	37 64 80 151	1	Rn-values	5	6	6%	4%	7%	0%
NCS	23 42.3 51.1 59.5 69.3 83.7	3	equally spaced	6	7	5%	3%	6%	0%
Bsp	25 46.9 59.5 75.5	3	equally spaced	7	35	4%	0%	3%	0%
Bsp	46.9 59.5 75.5	2	equally spaced	8	8	4%	3%	5%	0%
Bsp	25 46.9 59.5 75.5	2	equally spaced	9	19	4%	0%	4%	0%
NCS	16.67 29.33 42.3 51.1 59.5 69.3 83.7	3	equally spaced	10	22	3%	0%	2%	0%
Bsp	37 64 100	2	Rn-values	11	13	2%	2%	3%	0%
Bsp	37 64 151	2	Rn-values	12	14	2%	1%	2%	0%
Bsp	37 64 100	1	Rn-values	13	2	2%	9%	3%	0%
Bsp	46.9 59.5 75.5	3	equally spaced	14	26	2%	0%	2%	0%
Bsp	24 44.2 54.4 65.1 80.1	3	equally spaced	15	43	2%	0%	1%	0%
Bsp	37 64 80 151	2	Rn-values	16	27	2%	0%	1%	0%
Bsp	37 64 100 151	2	Rn-values	17	28	2%	0%	1%	0%
Bsp	51.1 69.3	2	equally spaced	18	3	1%	7%	2%	0%
Bsp	44.2 54.4 65.1 80.1	3	equally spaced	19	38	1%	0%	1%	0%
Bsp	37 64 100	3	Rn-values	20	31	1%	0%	1%	0%
Bsp	37 64 151	3	Rn-values	21	34	1%	0%	1%	0%
Bsp	51.1 69.3	3	equally spaced	22	17	1%	1%	1%	0%
Bsp	37 64 100 151	3	Rn-values	23	39	1%	0%	0%	0%
Bsp	37 64 80 151	3	Rn-values	24	40	1%	0%	0%	0%
NCS	18 32 46.9 59.5 75.5	3	equally spaced	25	11	0%	2%	1%	0%
Bsp	44.2 54.4 65.1 80.1	1	equally spaced	26	25	0%	0%	0%	0%
Bsp	24 44.2 54.4 65.1 80.1	1	equally spaced	27	37	0%	0%	0%	0%
NCS	37 64 100 151	3	Rn-values	28	4	0%	6%	0%	0%
Bsp	25 46.9 59.5 75.5	1	equally spaced	29	32	0%	0%	0%	0%
NCS	17.33 30.67 44.2 54.4 65.1 80.1	3	equally spaced	30	33	0%	0%	0%	0%
NCS	37 64 80 151	3	Rn-values	31	9	0%	2%	0%	0%
NCS	40.8 59.7 84.0 151	3	Harrell	32	10	0%	2%	0%	0%
NCS	24 44.2 54.4 65.1 80.1	3	equally spaced	33	24	0%	0%	0%	0%
NCS	25 46.9 59.5 75.5	3	equally spaced	34	15	0%	1%	0%	0%
Bsp	46.9 59.5 75.5	1	equally spaced	35	29	0%	0%	0%	0%
NCS	42.3 51.1 59.5 69.3 83.7	3	equally spaced	36	36	0%	0%	0%	0%
NCS	40.8 51.2 64.4 90.2	3	Harrell	37	20	0%	0%	0%	0%
NCS	44.2 54.4 65.1 80.1	3	equally spaced	38	21	0%	0%	0%	0%
NCS	46.9 59.5 75.5	3	equally spaced	39	12	0%	2%	0%	0%
Bsp	51.1 69.3	1	equally spaced	40	30	0%	0%	0%	0%
NCS	37 64 100	3	Rn-values	41	23	0%	0%	0%	0%
NCS	37 64 151	3	Rn-values	42	41	0%	0%	0%	0%
NCS	42.9 64.4 149	3	Harrell	43	42	0%	0%	0%	0%
NCS	51.1 69.3	3	equally spaced	44	44	0%	0%	0%	100%

Knot basis: Bsp B-splines; NCS natural cubic splines; Knot setting: “Rn-values” typical radon thresholds or fractions thereof (e.g. 50, 100, 100/3 or 200 Bq/m³), b) “equally spaced” knots in terms of exposure quantiles (e.g. then 25%-, 50%- and 75%-quartiles) or c) according a recommendation by “Harrell et al.” (Harrell 2001); AIC Akaike's information criteria – equal to BGIC(c=2, r=0); BIC Bayesian information criteria – equal to BGIC(c=1, r=1); BGIC(c,r) Bayesian-type generalized information criteria; model weight: best – lowest Herfindahl-Hirschman Index (HHI), worst – highest HHI; ΔAIC/ΔBIC/ΔBGIC difference in AIC/BIC/BGIC of models with and without a spline for sIR.

9 References

Andersen CE, Ulbak K, Damkjær A, Gravesen P (2001) Radon i danske boliger. Kortlægning af lands-, amts- og kommuneværdier. Sundhedsstyrelsen, Statens Institut for Strålehygiejne, København

Bleymüller J, Gehlert G, Gülicher H (2008) Statistik für Wirtschaftswissenschaftler, 15., überarb. Aufl. Vahlen, München

Bossew P (2010) Radon: exploring the log-normal mystery. J Environ Radioact 101:826–834. https://doi.org/10.1016/j.jenvrad.2010.05.005

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer-Verlag, New York

Cheng ES, Egger S, Hughes S, et al (2021) Systematic review and meta-analysis of residential radon and lung cancer in never-smokers. Eur Respir Rev 30:200230. https://doi.org/10.1183/16000617.0230-2020

Daraktchieva Z, Miles JCH, McColl N (2014) Radon, the lognormal distribution and deviation from it. J Radiol Prot 34:183–190. https://doi.org/10.1088/0952-4746/34/1/183

Darby S, Hill D, Auvinen A, et al (2005) Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ 330:223

Darby S, Hill D, Deo H, et al (2006) Residential radon and lung cancer--detailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14,208 persons without lung cancer from 13 epidemiologic studies in Europe. Scand J Work Environ Health 32 Suppl 1:1–83

Dobrzynski L, Fornalski KW, Reszczynska J (2018) Meta-analysis of thirty-two case-control and two ecological radon studies of lung cancer. J Radiat Res 59:149–163. https://doi.org/10.1093/jrr/rrx061

European Commission, Joint Research Centre, Čeliković I, et al (2019) Literature review of indoor radon surveys in Europe. Publications Office

Friedman H (2007) Das österreichische nationale Radonprojekt – Projekt Endbericht (ÖNARP). Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Wien, Österreich

G. Haquin, T. Riemer, Y. Shamai, et al (2012) Radon Survey of Israel

Grainger P, Shalla SH, Preece AW, Goodfellow SA (2000) Home radon levels and seasonal correction factors for the Isle of Man. Phys Med Biol 45:2247–52

Harrell FE (2001) Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. Springer Verlag, New York Berlin Heidelberg

Krewski D, Lubin JH, Zielinski JM, et al (2006) A combined analysis of North American case-control studies of residential radon and lung cancer. J Toxicol Environ Health A 69:533–597. https://doi.org/10.1080/15287390500260945

Krewski D, Lubin JH, Zielinski JM, et al (2005) Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. Epidemiology 16:137–145. https://doi.org/10.1097/01.ede.0000152522.80261.e3

Li C, Wang C, Yu J, et al (2020) Residential Radon and Histological Types of Lung Cancer: A Meta-Analysis of Case–Control Studies. Int J Environ 17:1457. https://doi.org/10.3390/ijerph17041457

Lorenzo-González M, Ruano-Ravina A, Torres-Durán M, et al (2019) Lung cancer and residential radon in never-smokers: A pooling study in the Northwest of Spain. Environ Res 172:713–718. https://doi.org/10.1016/j.envres.2019.03.011

Lu Z-H (2008) On an Expression of Generalized Information Criterion

Lubin JH, Wang ZY, Boice JD, et al (2004) Risk of lung cancer and residential radon in China: Pooled results of two studies. Int J Cancer 109:132–137. https://doi.org/10.1002/ijc.11683

Malinovsky G, Yarmoshenko I, Vasilyev A (2019) Meta-analysis of case-control studies on the relationship between lung cancer and indoor radon exposure. Radiat Environ Biophys 58:39–47. https://doi.org/10.1007/s00411-018-0770-5

Murphy P, Organo C (2008) A comparative study of lognormal, gamma and beta modelling in radon mapping with recommendations regarding bias, sample sizes and the treatment of outliers. J Radiol Prot 28:293–302. https://doi.org/10.1088/0952-4746/28/3/001

Rubin DB (2004) Multiple Imputation for Nonresponse in Surveys. Wiley & Sons Ltd

SAS (2013) Shared Concepts and Topics. In: SAS/STAT® 13.1 User's Guide. SAS Institute Inc., Cary, NC, USA, p 186

Tollefsen T, Cinelli G, Bossew P, et al (2014) From the European indoor radon map towards an atlas of natural radiation. Radiation Protection Dosimetry 162:129–134. https://doi.org/10.1093/rpd/ncu244

Tuia D, Kanevski M (2008) Indoor radon distribution in Switzerland: lognormality and Extreme Value Theory. Journal of Environmental Radioactivity 99:649–657. https://doi.org/10.1016/j.jenvrad.2007.09.004

Valmari T, Mäkeläinen I, Reisbacka H, Arvela H (2010) Suomen radonkartasto 2010 – Radonatlas över Finland 2010 – Radon Atlas of Finland 2010

Wang H, Zhang X, Zou G (2009) Frequentist model averaging estimation: a review. J Syst Sci Complex 22:732–748. https://doi.org/10.1007/s11424-009-9198-y

(2012) Cross-Canada Survey of Radon Concentrations in Homes - Final Report. Health Canada, Ottawa, Canada

(1993) EPA's Map of Radon Zones. United States Environmental Protection Agency; Office of Radiation and Indoor Air (ORIA)